DOCUMENT RESUME

ED 346 137 TM 018 457

AUTHOR Allen, Nancy L.; And Others
TITLE A Missing Data Approach to Estimating Distributions of Scores for Optional Test Sections.
SPONS AGENCY Educational Testing Service, Princeton, N.J.
PUB DATE Apr 92
NOTE 43p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 21-23, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Testing; *Estimation (Mathematics); Graphs; *Scaling; *Scores; Scoring; *Statistical Distributions; Test Content; *Test Format; Testing Problems; Testing Programs; Test Validity
IDENTIFIERS *Missing Data; Nonresponders; *Optional Test Questions

ABSTRACT

Many testing programs include a section of optional questions in addition to mandatory parts of a test. These optional parts of a test are not often truly parallel to one another, and groups of examinees selecting each optional test section are not equivalent to one another. This paper provides a general method based on missing-data methods for non-ignorable non-response (R.J.A. Little and D.B. Rubin, 1987) on an optional part of a test. The method is compared with other methods used to put optional test sections on comparable scales. If validity issues are resolved, this sensitivity analysis method could eventually provide a basis for the combination of scores based on optimal portions of a test with those based on common sections of a test. Eight tables and 31 graphs present results of the analysis. There is a 15-item list of references.
(Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED346137

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

NANCY L. ALLEN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A Missing Data Approach to Estimating Distributions of Scores for Optional Test Sections

Nancy L. Allen, Paul W. Holland, and Dorothy Thayer

Educational Testing Service

This work was supported by Educational Testing Service through the Program Research Planning Council. This paper was presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA, 1992.

TM018457

BEST COPY AVAILABLE

Abstract

Many testing programs include a section of optional questions in addition to mandatory parts of a test. For instance, some testing programs have an optional essay or performance portion of the test. These optional parts of a test are not often truly parallel to one another, and groups of examinees selecting each optional test section are not equivalent to one another. The purpose of this paper is to provide a general method based on missing-data methods for nonignorable nonresponse (Little & Rubin, 1987) to estimate distributions of scores on an optional part of a test. The method is compared with other methods used to put optional test sections on comparable scales. If validity issues are resolved, this method could eventually provide a basis for the combination of scores based on optional portions of a test with those based on common sections of a test.

A MISSING DATA APPROACH TO ESTIMATING DISTRIBUTIONS OF SCORES FOR OPTIONAL TEST SECTIONS

In some testing programs tests are administered that include optional tasks as well as mandatory sections. For instance, some tests include an optional essay or direct performance task. The fact that examinees must select the essay or task to which they respond means that the sample responding to any given prompt is not random, and may not represent the population of test-takers adequately. In such cases, the problem of finding valid and comparable scores no matter which optional test section is selected can be treated as a problem in which the samples of test-takers are not representative.

As a simple example, the design of the data from a test with two common sections and three optional essays is shown in Table 1. This design is similar in format to the more traditional testing setting where the same multiple-choice portion of a test is given in each administration, but the essay question is changed from administration to administration for security reasons. In this case, the examinee selects the time of administration. The design of the data in Table 1 is simple but can be extended to multiple administrations and more complex designs.

Insert Table 1 about here

In the future it is expected that other types of optional sections, such as portfolios or hands-on manipulation tasks, may be included more often in tests. Situations in which the

sampling of items is complex and sometimes not random is increasing due to the current movement toward more testing time devoted to examinee-constructed responses. As this trend develops it is clear that efforts to cover the entire domain of interest will not be successful due to time constraints. One popular way to circumvent this problem is to have examinees select examinee-constructed response tasks; in essence select their own form of the test based on their instructional background and interests and on their perception about the difficulty of the requested tasks. As indicated by Wainer, Wang, and Thissen (1991, p. 2), "The notion is that in this way the examinee is not placed at a disadvantage by an unfortunate choice of domain coverage by the test builder."

At present simple techniques making strong assumptions about the selection of optional portions of a test by examinees or about the characteristics of the optional test items are used to make the scores from optional sections comparable. Some of these methods assume that the examinees select optional questions in a random way or in a random way conditional on other information. Other methods recognize that the selection of optional test questions by an examinee may be a complex phenomenon, but have limiting assumptions about the characteristics of the optional items. For instance, the optional items may be assumed to be of equal difficulty. Most of these methods treat the missing information as ignorable in some way, when in fact the missing data are likely to be nonignorable (Little & Rubin, 1987).

A preliminary comparison of missing-data methods for cases of nonignorable nonresponse was completed by Allen and Holland (1989a). In that study, the distribution of scores for a single optional essay was estimated, given no extra information and, for the

ignorable case, given the results for the common multiple-choice portion of a test. The sensitivity analysis approach used in that study allows an examination of the effect of different assumptions about the nonresponse on a distribution of optional essay scores as if every student had selected that essay. The purpose of this paper is to put the sensitivity analysis approach in the perspective of previously proposed methods for adjusting for an optional essay, and presents advantages of the sensitivity analysis approach. In the future, these results can be built upon, estimating distributions for a single optional essay given the results of a common multiple-choice and a common essay section of a test under less stringent assumptions. Also the approach can be expanded to estimate multivariate distributions of test scores for more than one optional essay. Some issues involved in assigning scores to examinees are examined.

Although the methods described in this paper could be generalized to other settings, we specifically examine the case where a section of optional essays is presented to examinees, in addition to a common essay and/or a common multiple-choice section. The methods that are examined are applicable for any optional test section presented along with a common test section with an ordinal score. Our examples are based on data from the 1987 administration of a test having a format similar to that in Table 1, except that the design of the actual test has six optional essays, rather than the three in Table 1.

Previous Approaches

Approaches to this problem vary in their assumptions and in their complexities. At present simple techniques, making strong assumptions about the selection of optional portions of a test by examinees or about the characteristics of the optional essays are used

to make the scores from optional sections comparable. An approach to this problem that has been used in some testing programs is to calibrate the response readers across essays carefully and to ignore the fact that selection has taken place. This approach is based on the assumptions that the difference in the domains of the essays is not important and that the score levels on each essay have the same meaning.

Other approaches, such as those of Cowell (1972), Rosenbaum (1985) and Wainer, Wang and Thissen (1991), have assumptions that the examinees select optional questions in a random way or in a random way conditional on information other than the optional item score. Cowell (1972) proposed a method for adjusting optional essay scores based on regression methods related to the Pearson-Lawley adjustment to the correlation (Pearson, 1903; Lawley, 1943-44). Assumptions of the method include ignorable nonresponse conditional on the common portion of the test and strict selection on the basis of the common portion. Rosenbaum (1985) described an approach that treats the problem in a way that coincides with the approach of Little and Rubin (1987) to missing data. Rosenbaum's approach is based on propensity scores and makes the assumption of ignorable nonresponse given the propensity score. Wainer, Wang and Thissen (1991) approached the problem by modeling individual responses using item response functions and assume ignorable nonresponse given θ. To calibrate the item parameters for the common items and the selected essays they treat the examinees as two subpopulations, both administered the common items, but differing in the administration of the selected essays. They estimate an examinees overall proficiency using the expected a posteriori estimate of θ. They include a description of many of the issues that pertain to the problem.

Other methods recognize that the selection of optional test questions by an examinee may be a complex phenomenon, but have limiting assumptions about the characteristics of the optional items. For instance, in the first approach described, the optional items are assumed to be of equal difficulty. Livingston (1988) recognized that this assumption may not be appropriate in all cases, but also disagreed with the assumptions of Cowell's and Rosenbaum's methods because they result in adjusting essay scores most when the relationship between the common and optional portions of the test is weakest. Livingston weighted the amount of the adjustment by the correlation of the scores for the common and optional parts of the test.

A Sensitivity Analysis Approach

The approaches of Rubin (1987) and Tukey (Holland, 1986) are general approaches to missing data, designated in Wainer (1986) as mixture modeling and simplified selection modeling, respectively. These related approaches are developed here for application in the setting of an optional test section. We will begin with the simplest case where one optional essay is selected or not selected by an examinee and no information other than the essay score, if the essay was selected, is available. We will apply this simple approach to an example. We will then extend the approach to the case when more information is available. These approaches may be extended further in order to estimate distributions for all of the essays that may be selected by an examinee at once.

One-variable Case

Because terminology is important in understanding these approaches, we introduce some probabilities and distributions of interest here. The score on the optional portion of

the test is denoted Y. An indicator of whether the optional portion was selected or not is R, where R = 1 if and only if the value of Y is observed (i.e., not missing) and R = 0 if the value Y is missing. We adopt the following notation:

$p = P(R = 1) =$ response rate,

$g(y) = P(Y = y) =$ distribution of Y in the whole population,

$f_1(y) = P(Y = y|R = 1) =$ distribution of Y for those individuals for whom the value of Y is not missing,

$r(y) = P(R = 1|Y = y) =$ the selection function,

$f_0(y) = P(Y = y|R = 0) =$ distribution of Y for those individuals for whom the value of Y is missing.

The probability that an examinee chose the optional essay, p, and the distribution of the optional essay scores for examinees who selected the optional essay, $f_1(y)$, may be estimated from observed data. The distribution $f_0(y)$ is completely unknown (based on the observed Y-data) because it describes individuals whose Y-values are not observed.

Finally, g(y), the distribution of the optional essay scores for the whole population, will be the target quantity of interest in our study. We may also be interested in its mean value,

$$\sum_y y\, g(y)\ ,$$

or other moments of g. Information about the distribution g(y) could be used to transform scores on the optional portion of the test. Because of the possibility of missing data that are

not ignorable, g(y) cannot be estimated directly from the observed data. The distribution g(y) can be estimated using the relationship

$$g(y) = p\, f_1(y) + q\, f_0(y) \tag{1}$$

where $q = 1 - p$. This equation indicates that if the distribution of scores for the examinees who selected the optional essay and the distribution of scores for examinees who did not select the optional essay are combined properly we will get the distribution of interest, the distribution for all of the examinees.

Another important relationship between $f_1(y)$ and g(y) includes r(y):

$$g(y)\, r(y) = P(Y = y, R = 1) = p\, f_1(y). \tag{2}$$

It is useful to express r(y), the selection function, in its logit form via the definition

$$\lambda(y) = \ln[r(y)/(1 - r(y))] \tag{3}$$

so that

$$r(y) = \frac{e^{\lambda(y)}}{1 + e^{\lambda(y)}} \tag{4}$$

The logit, $\lambda(y)$, of the selection function is often useful in building models for r(y). For

example, in our analysis we will often consider the linear logit model for r(y) that is specified by

$$\lambda(y) = \lambda_0 + \lambda_1 y \tag{5}$$

so that

$$r(y) = \frac{e^{\lambda_0 + \lambda_1 y}}{1 + e^{\lambda_0 + \lambda_1 y}}, \tag{6}$$

where λ_0 and λ_1 are parameters that determine r(y).

Equations (1) and (2) and the definition (3) can be combined to obtain a formula that relates $\lambda(y)$ and $f_0(y)$:

$$f_0(y) = e^{-\lambda(y)} f_1(y) \Big/ \left[\sum_y e^{-\lambda(y)} f_1(y)\right]. \tag{7}$$

The proof of (7) is provided in Allen and Holland (in press). We will use equation (7) to specify choices of the distribution of scores for the examinees who did not select the optional essay, $f_0(y)$, in terms of $\lambda(y)$ and vice versa.

Now that we have described the notation and general relationships between quantities, our next goal is to discuss "sensitivity analysis." Sensitivity analyses are performed to illustrate how much conclusions about g(y) are affected by assumptions about the missing data.

From equation (1) we see that g(y) is determined in a simple way from p and $f_1(y)$ (both of which can be estimated from the observed data on Y and R) and $f_0(y)$ (about which

the observed data on Y and R have nothing to say at all). The most direct sensitivity analysis is to obtain estimates, $\hat{p}$ and $\hat{f}_1(y)$, from the observed data on R and Y, to consider various plausible choices of $f_0(y)$ and then to combine these three items via (1) to obtain plausible estimates of g(y), i.e.,

$$\hat{g}(y) = \hat{p}\,\hat{f}_1(y) + \hat{q}\,f_0(y)\,. \tag{8}$$

Glynn, Laird and Rubin (1986) call this direct approach "mixture modeling."

One problem with implementing this direct approach is that it may be difficult to know how to specify plausible choices of $f_0(y)$. An alternative is to first specify plausible choices for the selection function, r(y), or its logit, $\lambda(y)$, and then use (7) to generate the corresponding plausible choice of $f_0(y)$, i.e.,

$$f_0(y) = e^{-\lambda(y)}\hat{f}_1(y) \Big/ \left[\sum_y e^{-\lambda(y)}\hat{f}_1(y)\right]. \tag{9}$$

For example, we might try the simple linear logit (5) and vary λ_1 over some plausible range of values. This approach is implicit in what Tukey calls "simplified selection modeling" (Holland, 1986; Allen and Holland, 1989b). In this approach plausible choices of $\lambda(y)$ are specified and plausible estimates of g(y) are given by

$$\hat{g}(y) = \hat{p}\hat{f}_1(y) + \hat{q}\left\{ e^{-\lambda(y)}\hat{f}_1(y) \Big/ \left[\sum_y e^{-\lambda(y)}\hat{f}_1(y)\right] \right\}. \tag{10}$$

Note that, in the mixture approach [equation (8)], ignorable nonresponse is modeled by assuming that the distribution of essay scores is the same for those examinees who chose the essay [$f_1(y)$] and those who did not [$f_0(y)$]. In the ignorable case,

$$\hat{g}(y) = \hat{f}_1(y) = f_0(y).$$

In the selection function sensitivity analysis approach, ignorable nonresponse is modeled by assuming that the probability of selecting the optional essay does not depend on the essay score. This means that $r(y) = p$, which corresponds to $\lambda_1 = 0$ in the linear logit model.

In the linear logit model, the probability of the optional essay score, $Y = y$, being observed increases as the score, y, increases, when $\lambda_1 > 0$. When $\lambda_1 < 0$, the probability of observing the score decreases as the score increases. In many cases, this simple sort of quantitative information may be all that is available, making a linear logit sensitivity analysis model appealing.

Initially, in our example, we will emphasize the one-variable problem of estimating the distribution of scores that would have occurred had all 19,678 examinees who took the European History Advanced Placement in 1987 chosen to answer a single essay. Later, for one of the essays, essay six, we will take the objective test score into account to get an indication of which one-variable analysis might be most appropriate.

The data for the one-variable case consist of frequencies for each of the 15 possible scores for each optional essay. The total number of examinees selecting essay six, for instance, was 3,506. Therefore, the estimated response rate, $\hat{p}$, for this essay was .178. The sample sizes and test part means for examinees selecting each essay are in Table 2. Notice

that the mean scores for examinees selecting each essay are not ordered in the same way for the selected essay, the common essay and the objective portions of the test.

Insert Table 2 about here

The first step in the sensitivity analysis was to fit a log-linear model to the observed frequencies for the fifteen possible scores for each optional essay. For instance, for essay six, the model

$$f_1(y) = e^{\alpha + \beta_1 y + \beta_2 y^2 + \beta_3 y^3 + \beta_4 y^4 + \beta_5 y^5}, \quad y = 1, 2, \ldots, 15,$$

where α is the normalizing constant so that $\sum_y f_1(y) = 1$, was selected as the one most appropriate for the data. This model matches the first five moments of the observed data, including the mean (5.9), standard deviation (2.5) and skewness (.9). The models selected for each of the optional essays are listed in Table 3, along with information about the fit of the model to the data. The results of this step were estimated probabilities for the 15 score points in the distribution underlying the observed data for each of the essays.

Insert Table 3 about here

Next, various choices for $f_0(y)$ were made. Each of these choices was equivalent to a different linear logit model for r(y), so that equation (7) held. The first case we examined was the ignorable nonresponse case. Under this assumption, when $f_0(y)$ is combined with

$\hat{f}_1(y)$, the distribution does not change and the distribution of essay scores for all examinees, g(y), is the same as the estimated distribution for the observed data. Figure 1 contains the graph of the estimated distribution of the optional essay scores for all examinees, $\hat{g}(y)$, and the constant response function, $\hat{r}(y)$, for the sixth essay under different assumptions about λ_1.

Insert Figure 1 about here

In the ignorable nonresponse case,

$$\lambda_1 = 0$$

and

$$r(y) = \lambda_0 = \ln(\hat{p}/\hat{q}).$$

For the other cases, λ_1 was selected to be -0.2, -0.1, 0.1 and 0.2 and λ_0 contributed to the normalizing constant, α. This is equivalent to selecting

$$f_0(y) = e^{\alpha + (\beta_1 - \lambda_1)y + \beta_2 y^2 + \beta_3 y^3 + \beta_4 y^4 + \beta_5 y^5}, \; y = 1, 2, \ldots, 15 .$$

When λ_1 is negative the probability of selecting the sixth essay is larger for examinees who did or would have received lower scores on the essay. When λ_1 is positive the probability

of selecting the sixth essay is larger for examinees who did or would have received higher scores on the essay. Neither of these situations is necessarily unrealistic, given the information available in the one-variable case. It could be that all the other optional essays were very difficult, making it likely that examinees with lower ability should have selected the optional essay. If the examinees acted logically, then a negative value for λ_1 would be reasonable. It could be that only advanced students were introduced to the topics of the optional essay, making it unlikely that students of lower ability would select the optional essay. In this case a positive value for λ_1 would be reasonable. In order to draw conclusions about the reasonableness of these assumptions about λ_1 more information, such as that contained in the two-variable case, must be available. Tables 4, 5 and 6 contain the summary statistics for the estimated distribution of scores for all examinees, $\hat{g}(y)$, for each of the essays. When r(y) is decreasing (λ_1 negative) the mean score for all examinees is larger than when r(y) is increasing (λ_1 positive). The range of each of the summary statistics for different values of λ_1 vary by essay.

Insert Tables 4-6 about here

As seen in Figures 1 through 6, the results of the one-variable analyses provide several possible distributions for g(y) for each of the essays. The sensitivity analysis provided five (and could provide more) plausible solutions including the ignorable nonresponse case. As stated earlier, additional information about the most probable distribution of the essay scores for all examinees for the one-variable case can be gained from the objective test scores or information about other common test parts.

Insert Figures 2-6 about here

Two-variable Case

Now suppose that in addition to the optional essay score, Y, and the response indicator, R, there is a third variable, X = objective test score or other common test section score, that is fully observed, i.e., has no missing data. As stated earlier, in our example, this variable will contribute to information about the validity of the one-variable analyses. In most cases, a third variable will provide a basis for more complex analyses. Our notation is now easily expanded to condition everything on X.

$p(x) = P(R = 1|X = x) =$ response rate given X = x,

$g(y|x) = P(Y = y|X = x) =$ conditional distribution of Y given X in the whole population,

$f_1(y|x) = P(Y = y|R = 1, X = x) =$ conditional distribution of Y given X for those individuals for whom Y is not missing,

$r(y,x) = P(R = 1|Y = y, X = x) =$ the selection function,

$f_0(y|x) = P(Y = y|R = 0, X = x) =$ conditional distribution of Y given X for those individuals for whom Y is missing.

In addition, we need a notation for distributions of X. We have

$h(x) = P(X = x) =$ distribution of X in the whole population,

$h_1(x) = P(X = x|R = 1) =$ distribution of X for those individuals for whom Y is not missing,

$h_0(x) = P(X = x|R = 0) =$ distributions of X for those individuals for whom Y is

missing.

Comments parallel to those made in the one-variable case are now in order. First, p, p(x), $f_1(y|x)$, h(x), $h_1(x)$ and $h_0(x)$ may all be estimated from the observed data. Second, $f_0(y|x)$ is completely unknown (based on the observed Y- and X-data) because it describes the essay scores, Y, of individuals for whom scores are not observed. Third, g(y) is still the distribution of interest, but we will get at g(y) via the basic relationship

$$g(y) = \sum_x g(y|x)\, h(x) \,. \tag{11}$$

Hence g(y|x) is the intermediate target distribution. Because of missing data, g(y|x) cannot be estimated directly from the observed data. The generalization of (1) to the two-variable case is

$$g(y|x) = p(x)\, f_1(y|x) + q(x)\, f_0(y|x) \tag{12}$$

where q(x) = 1 - p(x).

As before, it is useful to express the selection function, r(y,x), in its logit form via the definition

$$\lambda(y,x) = \ln[r(y,x)/(1 - r(y,x))] \tag{13}$$

so that

$$r(y,x) = \frac{e^{\lambda(y,x)}}{1 + e^{\lambda(y,x)}} .$$

A logit function that is linear in y has the form

$$\lambda(y,x) = \lambda_0 + \lambda_1 y + \gamma(x) \tag{14}$$

for some function $\gamma(x)$.

We may combine (11), (12), and (13) to obtain a formula that relates $\lambda(y,x)$ and $f_0(y|x)$:

$$f_0(y|x) = e^{-\lambda(y,x)} f_1(y|x) \Big/ \left[\sum_y e^{-\lambda(y,x)} f_1(y|x)\right] . \tag{15}$$

We can use (15) to specify choices of $f_0(y|x)$ in terms of $\lambda(y,x)$ and vice versa.

One of the important concepts that arises when a fully observed variable, X, is available is that of ignorable nonresponse given X. This may be defined in several equivalent ways. Ignorable nonresponse given X parallels ignorable nonresponse in the one-variable case. The following conditions are equivalent ways of expressing ignorable nonresponse given X:

$$r(y,x) = p(x),$$

or

$$f_1(y|x) = f_0(y|x).$$

Sensitivity analysis can be extended in a simple way to the two-variable case. The key equation for mixture modeling is

$$\hat{g}(y|x) = \hat{p}(x)\,\hat{f}_1(y|x) + \hat{q}(x)\,f_0(y|x)\,. \tag{16}$$

The extension of the selection function sensitivity analysis to the two-variable case depends on the relationship

$$\hat{g}(y|x) = \hat{p}(x)\,\hat{f}_1(y|x) + \hat{q}(x)\left\{ e^{-\lambda(y,x)}\,\hat{f}_1(y|x) \Big/ \left[\sum_y e^{-\lambda(y,x)}\,\hat{f}_1(y|x)\right]\right\}\,. \tag{17}$$

In a way similar to the one-variable case, when using the mixture modeling approach, plausible choices of $f_0(y|x)$ are selected. In a selection function sensitivity analysis, plausible choices of the selection function, $r(y,x)$, or its logit, $\lambda(y,x)$, are specified.

We will consider the simplest two-variable case, the case of ignorable nonresponse given X. Ignorable nonresponse given X means that the conditional distributions of observed essay scores and unobserved essay scores on objective test scores are equal. This means that

$$\hat{g}(y|x) = \hat{f}_1(y|x).$$

In our example, the objective test scores ranged from 0 to 100. Therefore, $f_1(y|x)$ was

estimated by fitting a bivariate log-linear model to the 101 by 15 matrix of observed counts to get estimated probabilities for $f_1(y,x)$. This joint distribution was then divided by the distribution of X for those who selected the optional essay, $h_1(x)$, where these probabilities were estimated by the probabilities from the X marginal of $\hat{f}_1(y,x)$. The relationship

$$\hat{g}(y) = \sum_x \hat{g}(y|x)\, \hat{h}(x)$$

was then used to calculate $\hat{g}(y)$. The one-variable selection function, r(y), was estimated using the relationship in equation (2). The estimated values for $f_1(y)$ were those from the marginal Y distribution associated with $\hat{f}_1(y,x)$. The graphs of $\hat{r}(y)$ and $\hat{g}(y)$ for essay six are in Figure 7. The mean, standard deviation and skewness of $\hat{g}(y)$ were 6.5, 2.8 and 0.8, respectively. If nonresponse given the objective test score is ignorable, then the one-variable analysis best fitting the situation is the case where λ_1 was selected to be -0.1. In practice, if additional information from common test parts or even other relevant variables was available, a sensitivity analysis could be completed given that information. This example is meant to illustrate the fact that assumptions made conditional on other information determine the appropriate assumption about the missing data when the extra information is not used in the analysis.

Insert Figure 7 about here

Discussion

The major advantage of using the language introduced by Little and Rubin (1987)

to describe missing data is the ability to separate the information which is known from the information which is not known. Rosenbaum (1985) used the same terminology; however, other researchers have not, resulting in some confusion about the true assumptions that were made.

The major advantage of using a sensitivity analysis approach is that the effect of different assumptions about the nonresponse on the final distribution g(y) can be seen clearly. Previous methods, other than that of Livingston (1988), make a specific assumption that the nonresponse is ignorable given some other information. They do not allow for different theories about the nonresponse. Livingston acknowledges that the results of adjustments made on the basis of ignorable nonresponse may not be correct and tries to adjust for it. His alternative assumption, however, is that the essays are of equal difficulty, the same assumption that is made in practice in several testing programs. The sensitivity approach allows an examination of what the assumption of equal difficulty means in terms of the difference of results based on ignorable nonresponse. Equal difficulty of the optional essays means that no adjustment to the scores is made. In order to examine this situation each of the optional essays in our example was transformed to the scale of the common essay by matching means and standard deviations from the common essay and from the one-variable sensitivity analyses performed for each essay. A score of five for each essay was transformed to the common essay scale under each assumption about nonresponse. These transformed scores are in Table 7. If no adjustment was made and the score of five was reported for an examinee who selected essay two, the assumption of ignorable nonresponse would be very close to the assumption that is equivalent to saying that essay three was equally as difficult as the common essay. Five is very close to the value 4.9 reported as the

transformed score for the ignorable case for essay two. This is not the case for the other essays, however. If no adjustment was made and the score of five was reported for an examinee who selected essay one, the assumption that was being made would be more extreme than selecting λ_1 to be -0.2. For each of the essays, the score of five was below the mean. Table 8 shows the transformed values on the same scale for the score of ten, a score above the mean for each essay. Again, a difference in the assumption about nonresponse is implied for each essay when no adjustment is made to the scores.

Insert Tables 7 & 8 about here

Each of the pieces of previous research described above introduces the problem as a missing data problem, rather than a simple equating problem. This distinction is an important one. Rosenbaum's (1985) approach differs with the approach described here, in the probabilities that are actually estimated as well as in the assumption of ignorable nonresponse. For instance, the distribution $f_1(y)$ is fit in our analysis, in other words it is smoothed, while in Rosenbaum's analysis this distribution is estimated by sample proportions. Wainer, Wang and Thissen (1991) recognize that the assumption they are making is one of ignorable nonresponse given θ, and they provide an extensive discussion of the issues that are important in solving this problem.

Several important issues stem from applications of missing data methods to the problem of optional test sections. First, a problem that has been discussed under another guise is the problem that occurs when ignorable nonresponse methods are applied uniformly. As discussed by Livingston (1988), this problem is that the most adjustment in optional essay

score occurs when the correlation between the optional essay score and the common test section score is zero. The objection to this is based on an expectation that, if optional essays are expected to be of equal difficulty, then the most adjustment in the scores should occur when there is strong evidence to the contrary, in other words, when the correlation between the optional test scores and the common test section scores is high. This objection can be translated into the language of missing data by saying that the assumption of ignorable nonresponse given the common test section score is not reasonable; nonresponse to the optional essay depends on more than just the common test section score. The sensitivity analysis approach just described allows for this type of assumption.

Another important issue is the validity of the test scores. If individual diagnosis is not the only purpose for a test, test results must have a particular meaning for every examinee who takes the test. No matter which tasks an examinee selects as part of a test, the results across examinees will be used in the same way. For instance, if examinees are to be selected on the basis of the test score, the relationship between scores for different individuals must not depend on which tasks were selected by the examinee. If groups are to be compared on the basis of test score means, the distribution of scores must not be affected by which task examinees selected. In other words, an important concluding responsibility in the development of a test is to ensure the comparability of scores - to equate scores for different selected forms of a test. In tests with examinee-selected tasks, there are many issues that should be considered in the process of equating forms. One is the issue of whether the specialized knowledge required of the responses to the optional essays can be related to one another in any meaningful way. Another issue is the meaning of an adjusted essay score of, say 4.3, when the rubric used to score essays has descriptions

of only whole number scores. Finally, an important issue is the construct validity of score interpretation across multiple methods of measurement (Messick, 1991).

Conclusion

Equating techniques take advantage of either common populations or common test items to adjust scores on two forms of a test so that they are fair to examinees that take either form. The situation that occurs when the choice of an optional essay is given to the examinee is not amenable to equating techniques unless either (1) the common portion and selected portion of the test are similar to one another or (2) the distribution of the optional section scores is known for all examinees. The sensitivity analysis approach described here provides a method to obtain the distribution of optional section scores for all examinees, in a way that makes it clear what information can be estimated from the data and what information must be provided by assumption. Further, this method allows the practitioner to look at the effect of different assumptions about the nature of the nonresponse to the optional test sections. No other approach to the problem gives this opportunity to understand the nature of selection of optional test sections.

The approach does not allow the practitioner to yield responsibility for assuring the validity of the test scores for their stated purpose, however. Practitioners must demonstrate that final scores on a test have the same meaning no matter which optional test section was selected by the examinee.

Specific applications of the approach have shown that there is no final answer to the question, "What assumptions about nonresponse for optional test sections are appropriate?" The answer will vary from test to test and from optional test section to optional test section. By the very nature of the selection problem, final decisions about the nature of nonresponse

to optional test sections depend on one more piece of information. As seen in our example, an assumption of ignorable nonresponse given the additional information from the common multiple-choice section score determined the "correct" assumption about the nonresponse when only the optional essay score was available. An assumption of ignorable nonresponse given the common multiple-choice section score and the common essay score can determine the "correct" assumption about the nonresponse when only the optional essay score and the common multiple-choice section score are available. At some point, assumptions must be selected and stated clearly to complete an analysis taking nonresponse into account. The sensitivity analysis approach described here allows that to be done.

References

Allen, N. L. & Holland, P. W. (1989a). Approaches to nonignorable nonresponse with application to selection bias. A paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Allen, N. L. & Holland, P. W. (1989b). Exposing our ignorance: The only "solution" to selection bias. Journal of Educational Statistics, 14, 141-145.

Allen, N. L. & Holland, P. W. (in press). Approaches to nonignorable nonresponse with application to selection bias.

Cowell, W. R. (1972). A technique for equating essay question scores. Statistical Report SR-72-70. Princeton, NJ: Educational Testing Service.

Glynn, R. J., Laird, N. M. & Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In H. Wainer (Ed.), Drawing inferences from self-selected samples (pp. 115-142). New York: Springer-Verlag.

Holland, P. W. (1986). A comment on remarks by Rubin and Hartigan. In H. Wainer (Ed.), Drawing inferences from self-selected samples (pp. 149-151). New York: Springer-Verlag.

Lawley, D. N. (1943-44). A note on Karl Pearson's selection formulae. Royal Society of Edinburgh: Proceedings, Section A, 62, 28-30.

Little, R. J. A. & Rubin, D. B. (1987). Statistical analysis with missing data. New York: Wiley.

Livingston, S. A. (1988). Adjusting scores on examinations offering a choice of essay questions. Research Report RR-88-64. Princeton, NJ: Educational Testing Service.

Messick, S. (1991). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. Research Memorandum RM-91-1. Princeton, NJ: Educational Testing Service.

Pearson, K. (1903). Mathematical contributions to the theory of evolution -XI. On the influence of natural selection on the variability and correlation of organs. Philosophical Transactions of the Royal Society of London: Series A, 200, 1-66.

Rosenbaum, P. R. (1985). A generalization of direct adjustment, with an application to the scaling of essay scores. Program Statistics Research Technical Report No. 85-55. Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.

Wainer, H. (1986). Drawing inferences from self-selected samples. New York: Springer-Verlag.

Wainer, H., Wang, X. & Thissen, D. (1991). How well can we equate test forms that are constructed by examinees? Program Statistics Research Technical Report No. 91-15. Princeton, NJ: Educational Testing Service.

Table 1

Data from One Administration of a Test with a Common and an Optional Essay

Sub-population	Common multiple-choice items	Common essay	Selected essay 1	Selected essay 2	Selected essay 3
1	X	X	X		
2	X	X		X	
3	X	X			X

Table 2

Mean Scores on the Optional Essays and the Objective Portion of the Test for Examinees Who Chose Each Essay

	Essay Answered					
	1	2	3	4	5	6
Selected Essay	7.5	8.4	6.5	7.4	6.7	5.9
Common Essay	8.2	8.5	8.5	8.5	8.5	8.1
Objective Portion	51.8	57.7	53.3	52.0	51.7	42.7
Number of Examinees Selecting Essay	6261	2642	3104	1996	2139	3506

Table 3

Models for $\hat{f}_1(y)$

Essay	Number of Overall Moments	Selected Groups of Points Fit	Number of Moments for Selected Points	Likelihood Ratio	DF
1	5	7	none	42.31	8
2	3	6,7,8,9,10	2	17.79	8
3	6	none	none	23.56	8
4	2	6,7,13,14,15	2	15.45	9
5	2	8,9,10,11,12, 13,14,15	1	7.44	10
6	5	none	none	25.58	9

Table 4

Means for $\hat{g}(y)$

Essay	$\lambda_1 = -0.2$	$\lambda_1 = -0.1$	Ignorable Case $\lambda_1 = 0.0$	$\lambda_1 = 0.1$	$\lambda_1 = 0.2$
1	8.5	8.0	7.5	7.1	6.8
2	9.3	8.9	8.4	8.0	7.5
3	7.8	7.1	6.5	6.0	5.5
4	8.6	8.0	7.4	6.8	6.3
5	7.7	7.2	6.7	6.2	5.7
6	7.2	6.5	5.9	5.4	5.0

Table 5

Standard Deviations for $\hat{g}(y)$

Essay	$\lambda_1 = -0.2$	$\lambda_1 = -0.1$	Ignorable Case $\lambda_1 = 0.0$	$\lambda_1 = 0.1$	$\lambda_1 = 0.2$
1	2.7	2.6	2.5	2.4	2.4
2	2.3	2.3	2.3	2.3	2.4
3	2.9	2.8	2.6	2.5	2.4
4	2.8	2.6	2.5	2.4	2.4
5	2.5	2.4	2.4	2.3	2.3
6	3.1	2.8	2.5	2.3	2.2

Table 6

Skewness for $\hat{g}(y)$

Essay	$\lambda_1 = -0.2$	$\lambda_1 = -0.1$	Ignorable Case $\lambda_1 = 0.0$	$\lambda_1 = 0.1$	$\lambda_1 = 0.2$
1	.3	.4	.4	.4	.4
2	0.0	0.0	0.0	-.1	-.2
3	.3	.5	.5	.6	.6
4	.3	.4	.4	.4	.3
5	.3	.3	.2	.1	.1
6	.6	.8	.9	.9	.9

Table 7

Scores of Five on Each Essay Transformed to the Common Essay Scale by Matching Two Moments

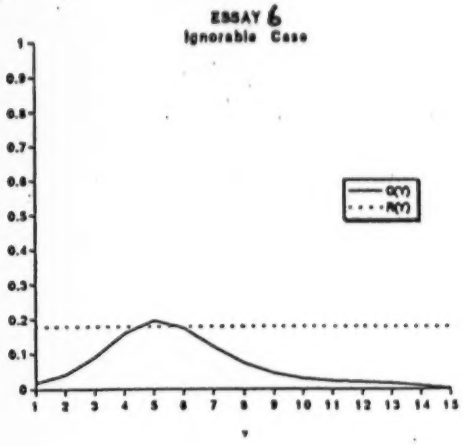
Essay	$\lambda_1 = -0.2$	$\lambda_1 = -0.1$	Ignorable Case $\lambda_1 = 0.0$	$\lambda_1 = 0.1$	$\lambda_1 = 0.2$
1	5.4	5.7	6.0	6.3	6.6
2	4.1	4.5	4.9	5.3	5.9
3	6.1	6.6	6.9	7.3	7.8
4	5.4	5.7	6.1	6.6	7.0
5	5.8	6.2	6.7	7.1	7.6
6	6.7	7.0	7.4	7.8	8.2

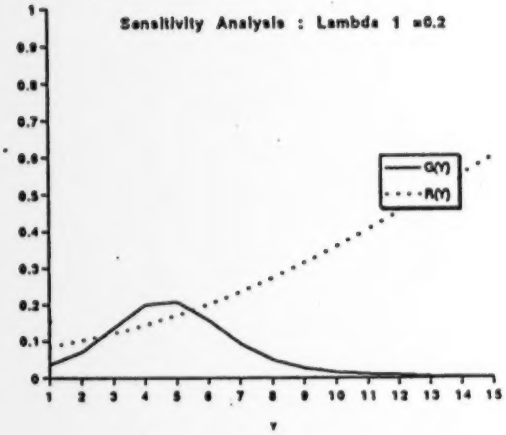
Table 8

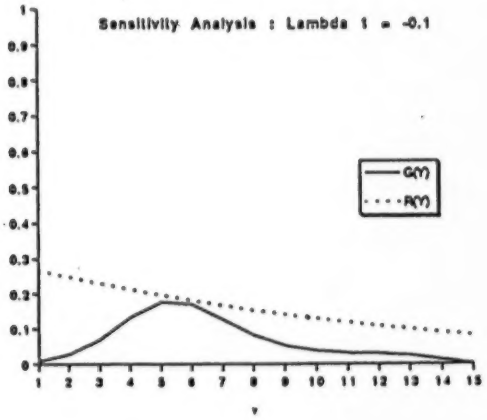
Scores of Ten on Each Essay Transformed to the Common Essay Scale by Matching Two Moments

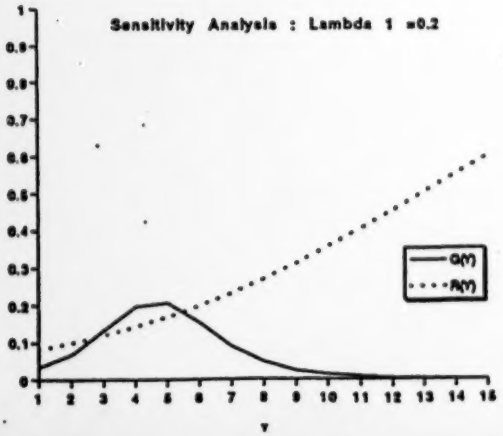
Essay	$\lambda_1 = -0.2$	$\lambda_1 = -0.1$	Ignorable Case $\lambda_1 = 0.0$	$\lambda_1 = 0.1$	$\lambda_1 = 0.2$
1	9.5	9.9	10.5	10.9	11.2
2	8.9	9.3	9.8	10.2	10.5
3	9.9	10.5	11.2	11.8	12.4
4	9.3	9.9	10.5	11.2	11.7
5	10.3	10.8	11.3	11.9	12.4
6	10.2	11.0	11.9	12.7	13.3

Figure 1


ESSAY 6
Ignorable Case
Q(Y)
R(Y)
Y


Sensitivity Analysis : Lambda 1 =0.2
Q(Y)
R(Y)
Y


Sensitivity Analysis : Lambda 1 = -0.1
Q(Y)
R(Y)
Y


Sensitivity Analysis : Lambda 1 =0.2
Q(Y)
R(Y)
Y

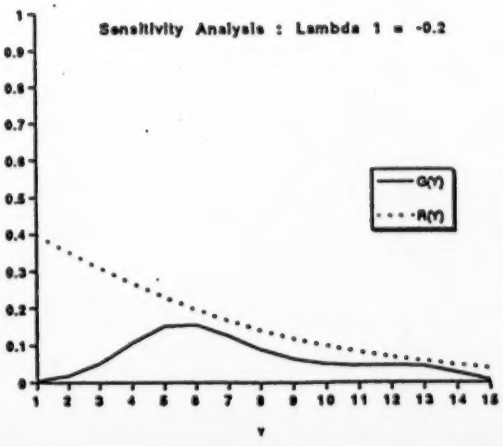
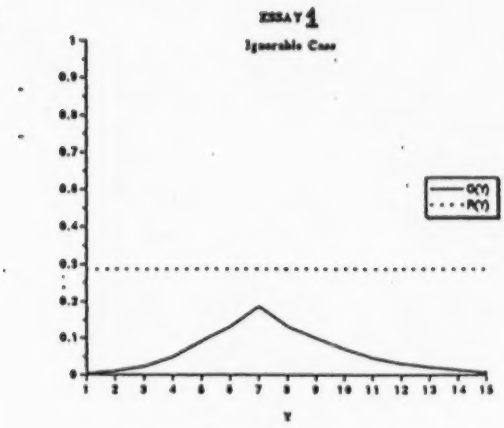
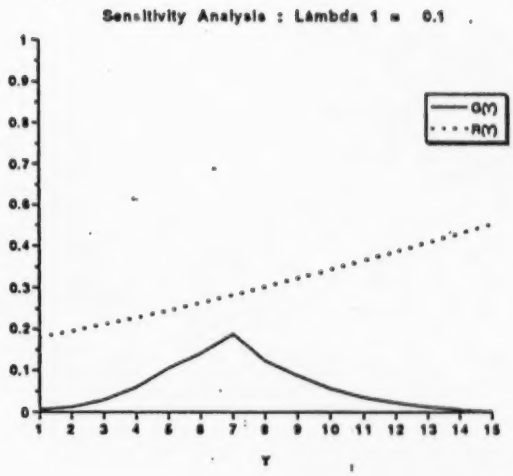
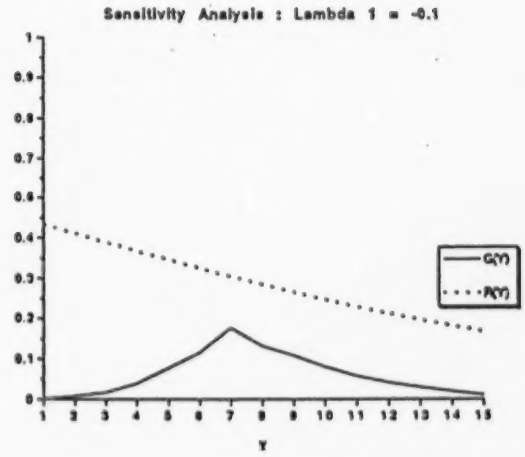
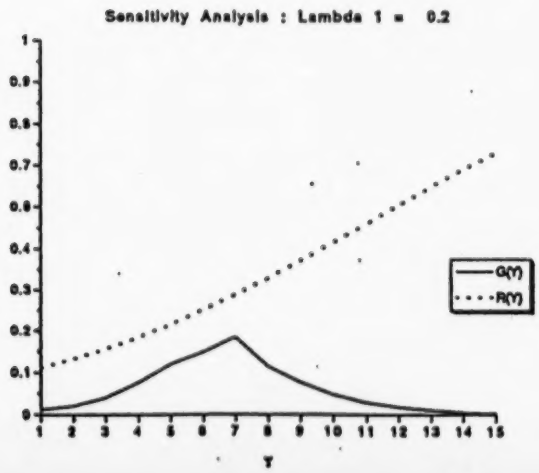

Sensitivity Analysis : Lambda 1 = -0.2
Q(Y)
R(Y)
Y

Figure 2


ESSAY 1
Ignorable Case
G(Y)
R(Y)
Y


Sensitivity Analysis : Lambda 1 = 0.1
G(Y)
R(Y)
Y


Sensitivity Analysis : Lambda 1 = -0.1
G(Y)
R(Y)
Y


Sensitivity Analysis : Lambda 1 = 0.2
G(Y)
R(Y)
Y

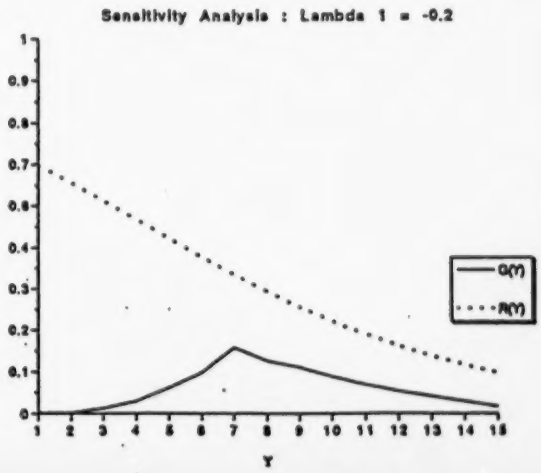
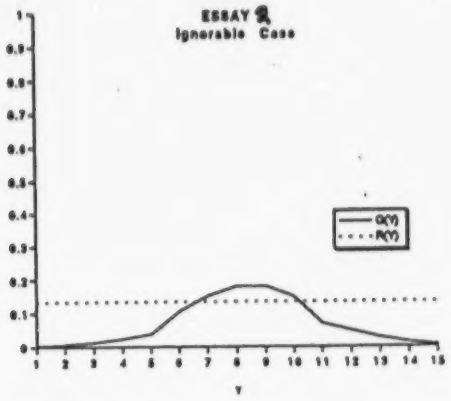
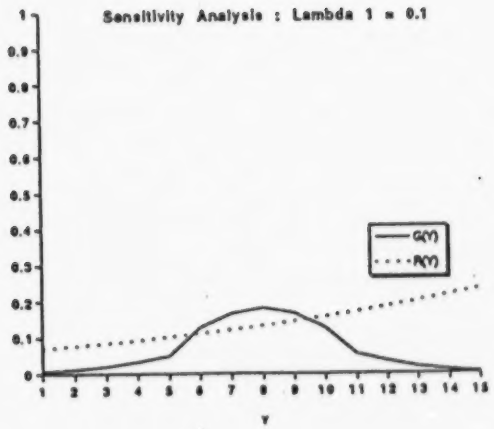
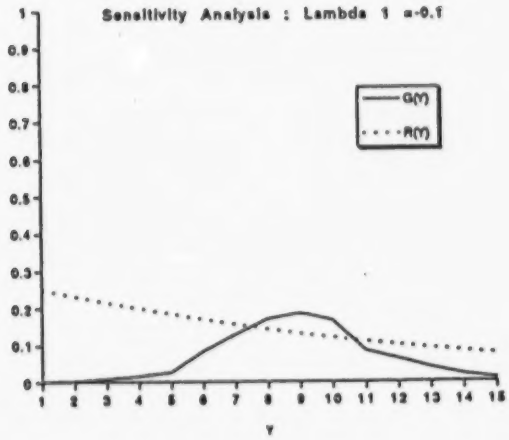
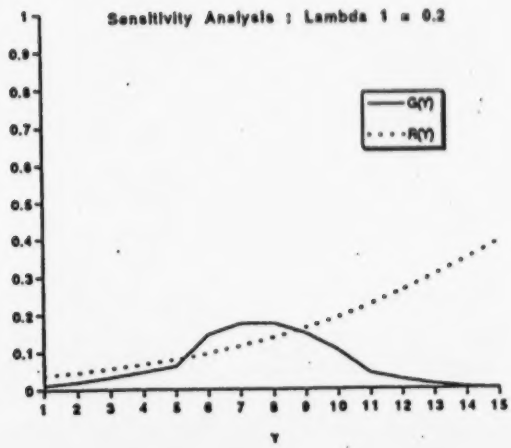

Sensitivity Analysis : Lambda 1 = -0.2
G(Y)
R(Y)
Y

Figure 3


ESSAY
Ignorable Case
G(Y)
R(Y)
Y


Sensitivity Analysis : Lambda 1 = 0.1
G(Y)
R(Y)
Y


Sensitivity Analysis : Lambda 1 =-0.1
G(Y)
R(Y)
Y


Sensitivity Analysis : Lambda 1 = 0.2
G(Y)
R(Y)
Y

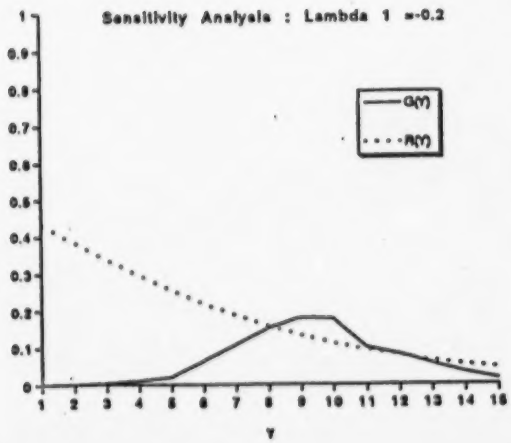
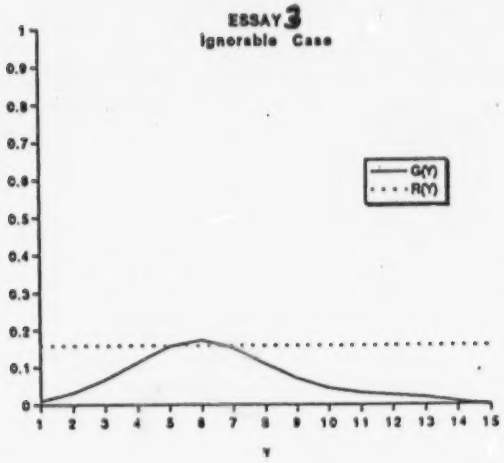
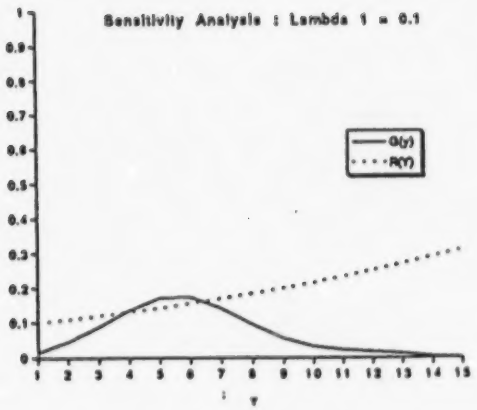
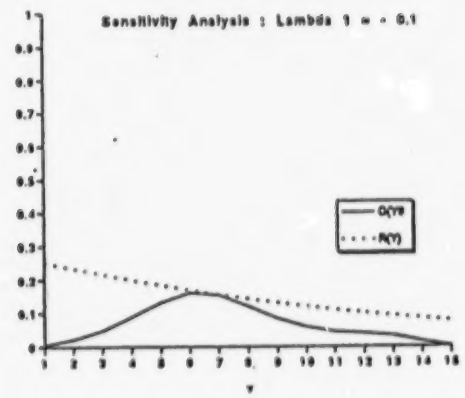
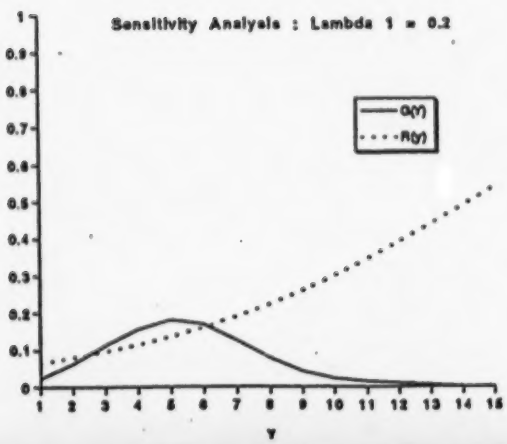

Sensitivity Analysis : Lambda 1 =-0.2
G(Y)
R(Y)
Y

Figure 4


ESSAY 3
Ignorable Case
G(Y)
R(Y)
Y


Sensitivity Analysis : Lambda 1 = 0.1
G(Y)
R(Y)
Y


Sensitivity Analysis : Lambda 1 = - 0.1
G(Y)
R(Y)
Y


Sensitivity Analysis : Lambda 1 = 0.2
G(Y)
R(Y)
Y

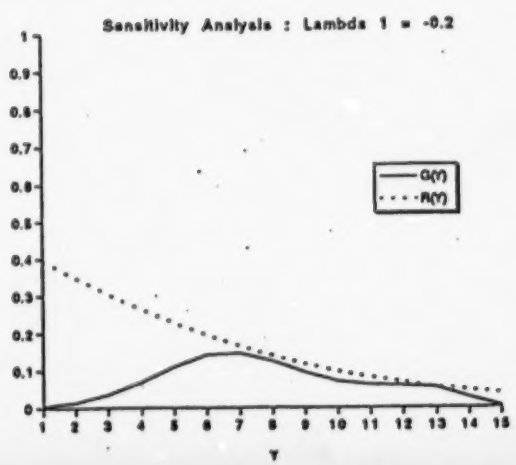
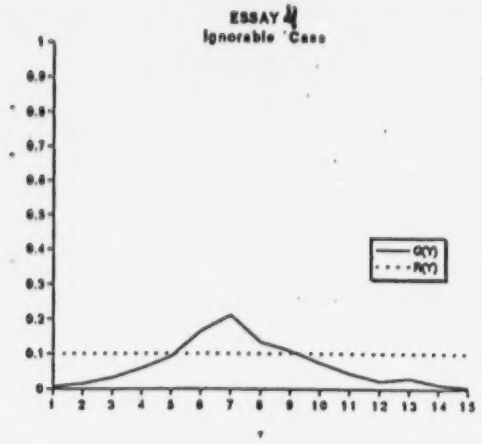
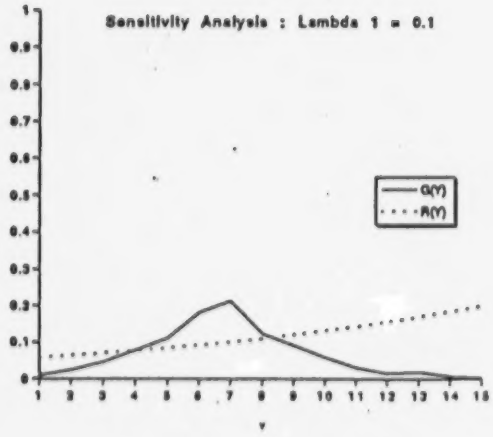
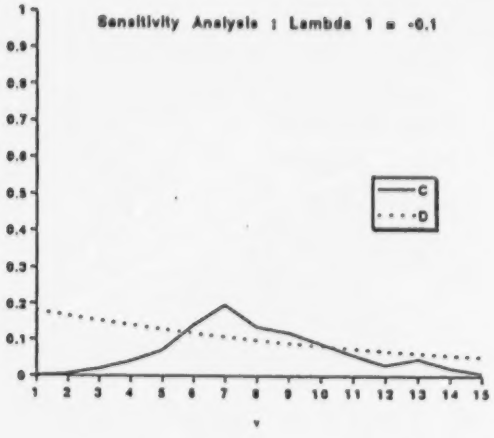
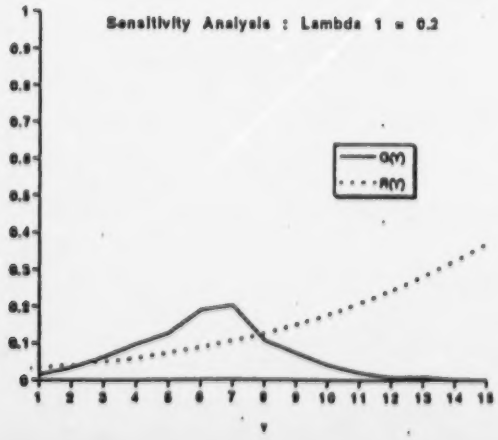

Sensitivity Analysis : Lambda 1 = -0.2
G(Y)
R(Y)
Y

Figure 5


ESSAY 4
Ignorable Case
G(Y)
R(Y)
Y


Sensitivity Analysis : Lambda 1 = 0.1
G(Y)
R(Y)
Y


Sensitivity Analysis : Lambda 1 = -0.1
C
D
Y


Sensitivity Analysis : Lambda 1 = 0.2
G(Y)
R(Y)
Y

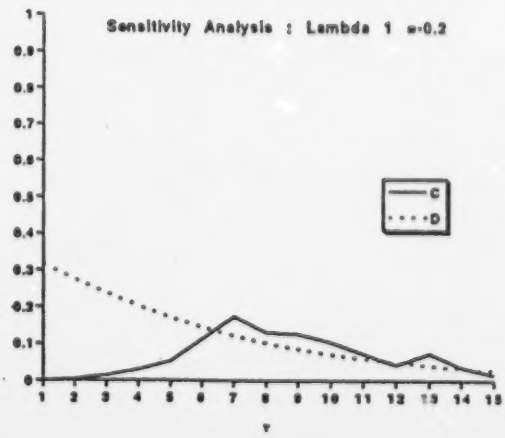

Sensitivity Analysis : Lambda 1 =-0.2
C
D
Y

Figure 6

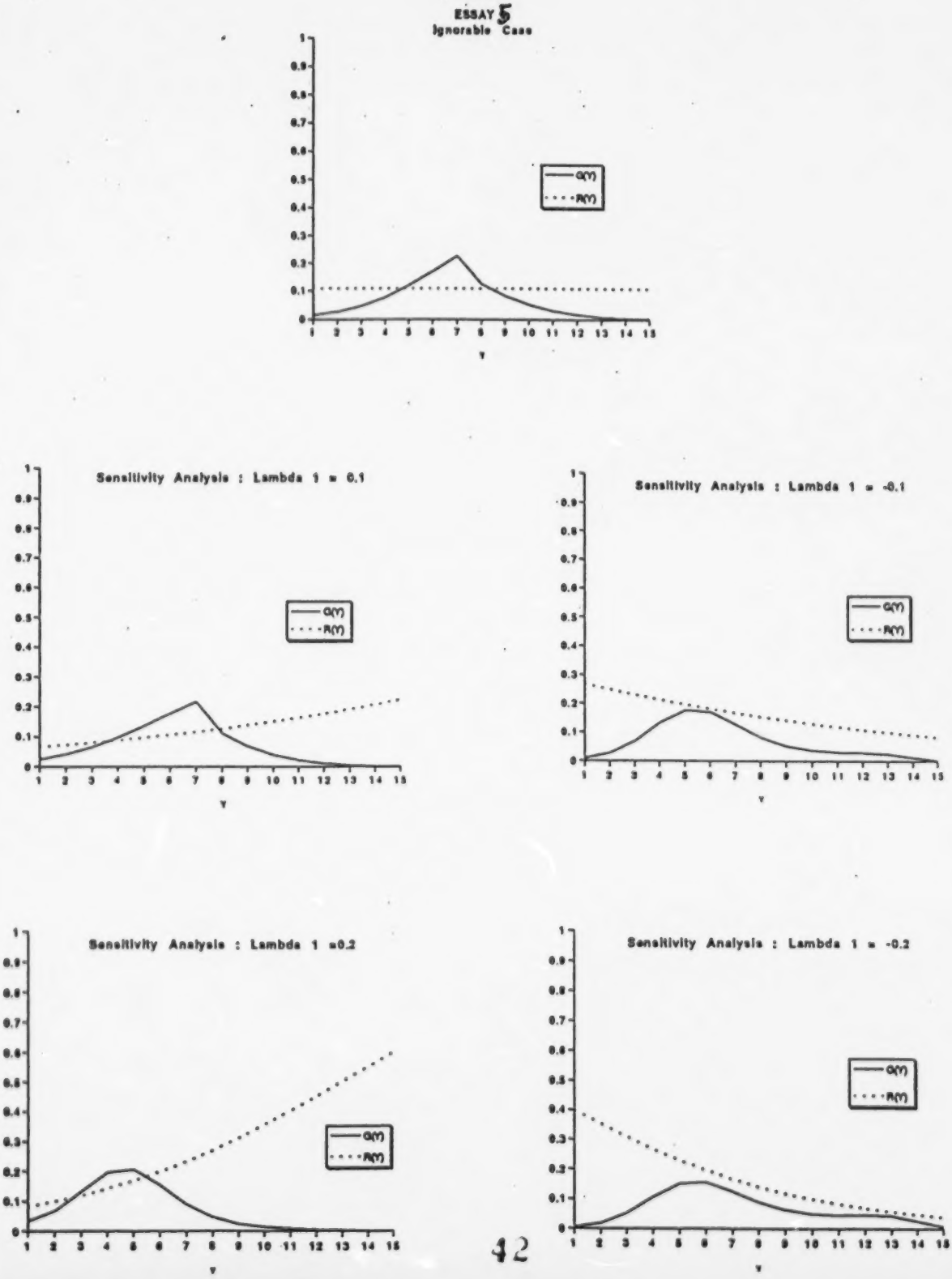
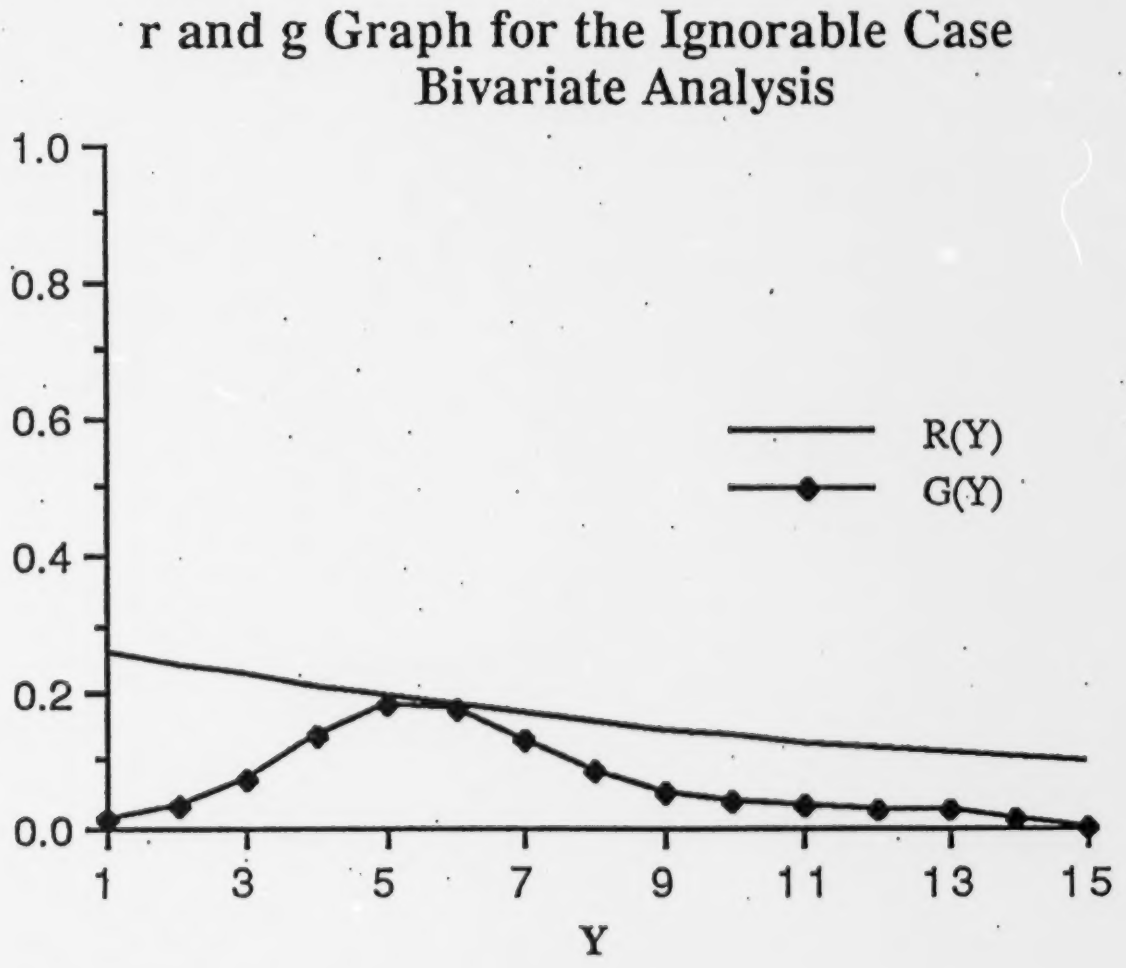

ESSAY 5
Ignorable Case
Q(Y)
R(Y)
Y
Sensitivity Analysis : Lambda 1 = 0.1
Sensitivity Analysis : Lambda 1 = -0.1
Sensitivity Analysis : Lambda 1 =0.2
Sensitivity Analysis : Lambda 1 = -0.2

Figure 7


r and g Graph for the Ignorable Case
Bivariate Analysis
1.0
0.8
0.6
0.4
0.2
0.0
R(Y)
G(Y)
1
3
5
7
9
11
13
15
Y